

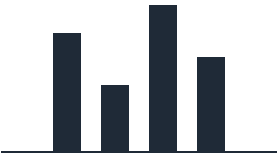
AP® Statistics

1A

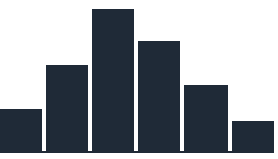
Exploring One-Variable Data



DOTPLOT



BAR CHART



HISTOGRAM

1		2	5	8	
2		0	3	4	7
3		1	6	6	9
4		2	5		
5		1			

STEM-AND-LEAF

Guided Notes Packet

DSAHS 2026-2027 · MR. LEVY

_____ is the science and art of _____ , analyzing, and drawing _____ from data.

A DATA SET: FAVORITE MOVIES IN ONE CLASS

STUDENT	FAVORITE MOVIE	GENRE	PLATFORM	RATING (/10)
Alice	Inception	Sci-Fi	Netflix	9
Bob	The Dark Knight	Action	Hulu	10
Charlie	The Shawshank Redemption	Drama	Amazon Prime	8
Diana	Parasite	Thriller	Netflix	9
Edward	Toy Story	Animation	Disney+	7
Fiona	Titanic	Romance	Hulu	8
George	Avengers: Endgame	Superhero	Disney+	9
Hannah	Get Out	Horror	HBO Max	8
Isaac	Forrest Gump	Drama	Amazon Prime	9
Julia	The Matrix	Sci-Fi	Netflix	10

Keep this table handy. We come back to it later in the section.

Smelling Parkinson's. Parkinson's disease is a progressive neurological disorder that primarily affects _____. Joy Milne claims she can detect it by smell. Answer as we watch:

- a) Why would it be important to know that someone can smell Parkinson's disease?

- b) How many correct decisions (out of 12) would it take to convince you that Joy really could smell Parkinson's?

- c) Which is more believable: Joy is a lucky guesser, or Joy can really smell Parkinson's? Explain.

A statistical study starts with an _____ question. A valid statistical investigative question is one based on data that _____ .

VALID ✓
"Can Joy Milne really smell Parkinson's?" Some shirts were identified correctly, others not. The data vary.

NOT VALID ✕
"What is the world record for the 100-meter dash?" A single data value answers it. Nothing varies.

Write your own valid investigative question about students at our school:

Population

The _____ group of individuals we want information about.

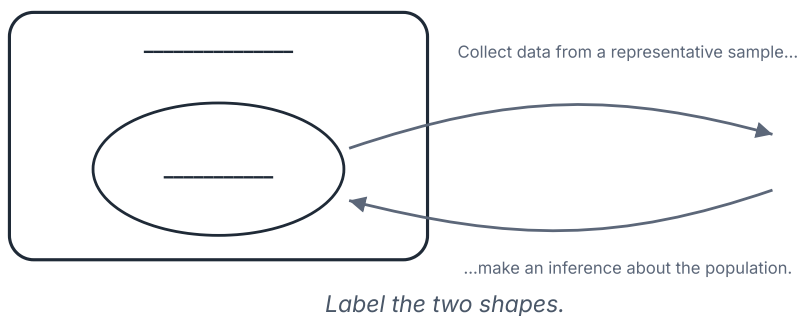
Census

Collects data from _____ individual in the population.

Sample

A _____ of individuals in the population from which we actually collect data.

Why not just take a census every time?



Understanding Check. Identify the population and the sample in each situation.

- a) To estimate this week's average ER wait time, hospital staff time 50 patients from arrival to treatment.
-
-

- b) A company selects 20 lifeguard shifts from the next 2 months' schedule for unannounced skills tests.
-
-

Fish in Lake Malawi. Imagine we are interested in the mean length of fish in Lake Malawi (Tanzania). Would it be reasonable to compute it? What might be a reasonable way to estimate it?

Any number that summarizes data from a _____ is a statistic. Any number that summarizes data for a _____ is a parameter. Examples of both: minimum, maximum, median, mean, range, standard deviation, a percent above some value.

Parameter

A number that describes some characteristic of a _____.

Statistic

A number that describes some characteristic of a _____.

We use a sample _____ to estimate a population _____.

Notation: n = _____ size; N = _____ size.

An _____ is an object described in a set of data.

Observational units can be people, animals, or things.

A _____ is an attribute that can take different values for different observational units.

CATEGORICAL VARIABLE

Assigns _____ that place each observational unit into a particular group, called a category.

QUANTITATIVE VARIABLE

Takes _____ values that are quantities: counts or measurements.

Go back to the movie table in 1A.1. Star the column that identifies each observational unit, then write a C or a Q above each of the other column headers.

Zip codes are made of digits. Why is a zip code *not* a quantitative variable?

VARIABLE SORT

Annual income of households in dollars	Number of siblings each student has
Average daily temperature in degrees Celsius	Preferred genre of music
Blood type	Shirt sizes
Car brand owned by students	Time to complete a math quiz in minutes
Distance traveled to reach Soffer in miles	Type of transportation to school
Grade level (Fr, So, Jr, Sr)	Weight of backpacks carried by students
Height of students in centimeters	Zip codes of Soffer staff households
Number of pages in a textbook	

Separate the list into a categorical bucket and a quantitative bucket by annotating it: write C or Q next to each variable.

Let's Practice.

AIRPORT NAME	ELEV. (FT)	MUNICIPALITY	COUNTRY	RADIO FREQ. (MHZ)	RUNWAY LENGTH (FT)	SURFACE
Los Angeles Int'l	125	Los Angeles	United States	36.07	8925	Concrete
Kempegowda Int'l	3000	Bangalore	India	121.25	13,123	Asphalt
Sendai	15	Natori	Japan	120.4	9842	Asphalt
Warsaw Chopin	362	Warsaw	Poland	128.8	9186	Concrete
Brisbane Int'l	13	Brisbane	Australia	124.35	10,827	Asphalt
Jorge Newbery	18	Buenos Aires	Argentina	120.6	7710	Concrete
Seychelles Int'l	10	Mahe Island	Seychelles	119.7	9800	Concrete
Sofia	1742	Sofia	Bulgaria	123.7	11,811	Asphalt
Hewanorra Int'l	14	Vieux Fort	Saint Lucia	119.8	9003	Asphalt
La Guardia	21	New York	United States	118	7000	Asphalt

a) Identify the observational units and variables in this data set.

b) Classify each variable as categorical or quantitative.

The Statistical Problem-Solving Process. Every statistical study follows the same four-step rhythm. We will use it all year.

- ① _____ **questions:** pose a valid investigative question about a population.
- ② _____ **data:** design and carry out a plan for gathering data (a sample, an experiment).
- ③ _____ **data:** organize, display, and summarize what you collected.
- ④ _____ **results:** answer the investigative question in context.

The _____ of a variable tells us what values the variable takes and how _____ it takes those values. A variable generally takes values that vary; we are interested in the pattern of that variation.

A _____ table shows the *number* of observational units in each category. A _____ table shows the *proportion* in each category. Percentages, relative frequencies, and ratios all carry the same information as proportions.

CENSUS AT SCHOOL: 10 RANDOMLY SELECTED CANADIAN STUDENTS

PROVINCE	GENDER	LANGUAGES SPOKEN	HANDEDNESS	HEIGHT (CM)	WRIST CIRC. (MM)	PREFERRED COMMUNICATION
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

Build both tables for the variable "Province" from the data above. Include a Total row.

FREQUENCY TABLE

PROVINCE	FREQUENCY
TOTAL	

RELATIVE FREQUENCY TABLE

PROVINCE	REL. FREQ.
TOTAL	

To display the distribution of categorical data, make a _____ or a _____.

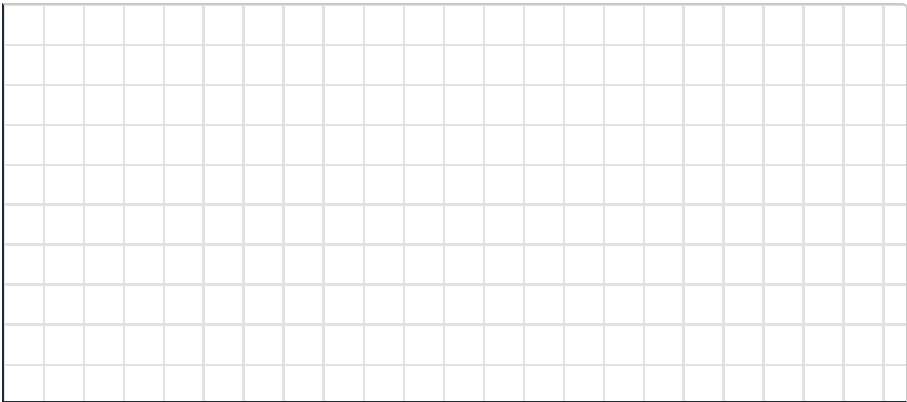
A **bar graph** shows each category as a bar. The heights of the bars show the category _____ or relative frequencies.

How to make a bar graph:

- 1) Draw and _____ the axes. Indicate whether the vertical axis measures frequency or relative frequency.
- 2) Write category names at _____ spaced intervals under the horizontal axis.
- 3) Draw bars above the category names, all with _____ width.

Practice: Bar Graph. A class voted on the nicest fruit. Draw the relative frequency bar graph:

FRUIT	APPLE	ORANGE	BANANA	KIWIFRUIT	BLUEBERRY	GRAPES
Rel. freq.	0.24	0.21	0.07	0.17	0.28	0.03



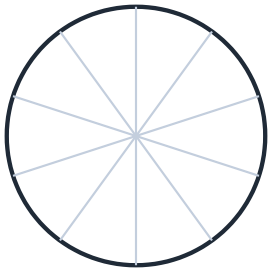
LABEL BOTH AXES AND TITLE THE GRAPH

A **pie chart** displays each category as a slice of the pie. The _____ of each slice is proportional to the category frequency or relative frequency.

Use a pie chart to emphasize each category's relation to the _____. A pie chart must include **ALL** observational units; add an " _____ " category if needed.

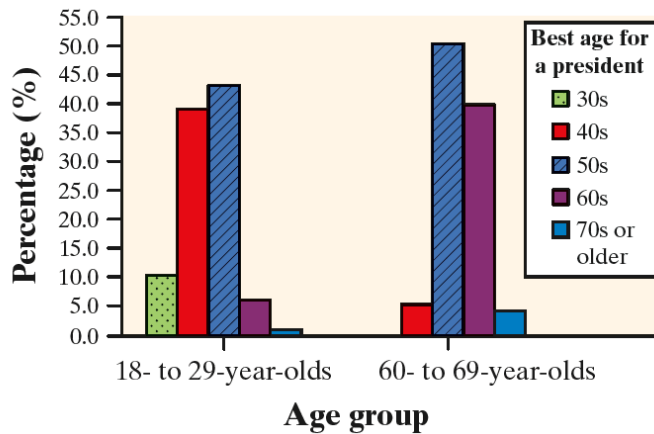
Practice: Pie Chart. Ten students named their preferred communication method. Build the pie chart (each sector = one student):

PREFERRED METHOD	FREQUENCY
Cell phone	2
Facebook	1
In person	3
Text messaging	4



A _____ bar chart displays the distribution of a categorical variable in two or more groups: a separate bar for each group in every category. Always include a _____ !

Pew Research asked U.S. adults the best age for a president. Compare the distributions for 18-to-29-year-olds and 60-to-69-year-olds:



Reaching Conclusions from Graphs. For the graph shown in class:

- What is misleading about the graph?
- How could the graph be changed to reflect the data more accurately?
- Why would someone make a graph that is misleading in this way?

Spotting a misleading graph; watch for:

Pictographs. Bar graphs are a bit dull to look at. It is tempting to replace the bars with pictures or 3-D effects to make the graph more interesting.

Don't do it! The eye responds to _____, so scaled pictures exaggerate differences.

A **dotplot** shows each data value as a _____ above its location on a number line. It is the simplest graph for a quantitative variable and easy to make by hand for small data sets.

Movie lengths (minutes): 110 115 110 130 117 142 120 121 125 125 126
127 131 131 130 133 130 119 128 133 144 106 143 148 163 131 134

- Make a dotplot of these data.

MOVIE LENGTH (MINUTES)

- b) A movie over two hours is considered long. Can the graph justify the claim that more than 75% of these movies are long? Explain.

A **stem-and-leaf plot** splits each value into a _____ (the leftmost digits) and a _____ (the final digit). Both are ordered least to greatest, and a key is required.

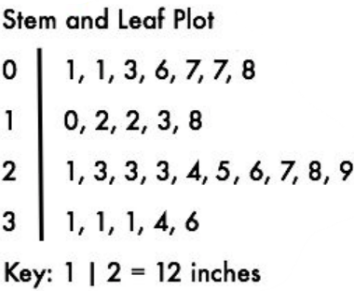
Pairs of shoes owned by 20 female AP Statistics students: 50 26 26 31
57 19 24 22 23 38 13 50 13 34 23 30 49 13 15 51

Construct the stemplot, then write the key.

Key: _____

When data values are bunched up, get a better picture by _____ stems. Two distributions of the same variable can be compared with a _____ stemplot using common stems.

Practice: Stemplots. Create a split-stem stemplot from the data in this stemplot:



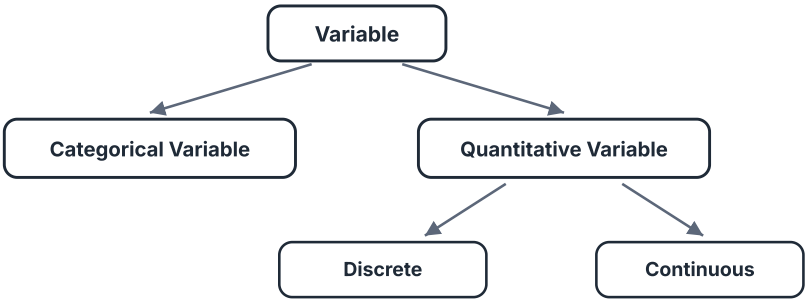
What is the range of the data? _____

A **histogram** shows each _____ of values as a bar. The heights show the frequencies or relative frequencies of values in each interval.

Three ways a histogram differs from a bar graph:

- 1) _____
- 2) _____
- 3) _____

Caution: use percents instead of counts when comparing groups of different sizes, and changing the bin width can change the histogram's appearance.



Discrete Variable

Takes a _____ number of numeric values: specific and separate (finite or countably infinite). Ex: number of doctor visits, number of Skittles in a bag.

Continuous Variable

A _____ quantity that can take infinitely many values in an interval; the values cannot be counted. Ex: height, time.

Let's Practice. Mark each variable C (categorical), DQ (discrete quantitative), or CQ (continuous quantitative):

____ Number of Marvel movies watched	____ Time spent commuting to school
____ Favorite fast-food restaurant	____ Steps taken between 1st and 2nd period
____ The weight of an elephant	____ Favorite social media platform
____ Tickets sold to different music festivals	____ Amount of lemonade in a jug at Chick-Fil-A

Describing distributions. Don't forget your **SOCS!** When describing a quantitative distribution, address all four, always in context:

S _____ : symmetric or skewed? peaks, clusters, gaps

O _____ : any values that stand apart?

C _____ : a typical or middle value

S _____ : the variability of the data values

Shape. A distribution is roughly _____ if the right side of the graph is approximately a mirror image of the left side. It is skewed to the _____ if the right side is much longer, and skewed to the _____ if the left side is much longer.

Sketch the three shapes as we discuss them:



SYMMETRIC



SKEWED RIGHT



SKEWED LEFT

Center

One value that best describes _____ of the data.

Outlier

A value that lies _____ from the rest of the data. Always think of possible reasons for outliers.

Gap

A region of a distribution between two data values where there are _____ observed data.

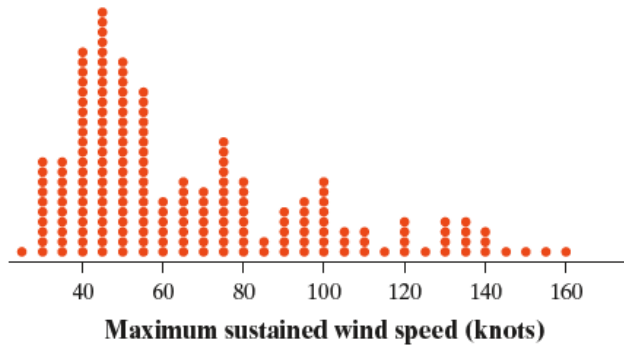
Cluster

Concentrations of data usually separated by _____.

Variability (Spread)

The _____ the data values fall in, and where the majority of the data lies.

Cyclone speeds. The dotplot shows maximum sustained wind speed (knots) for 193 Atlantic-basin tropical cyclones over 11 years. Describe the distribution.



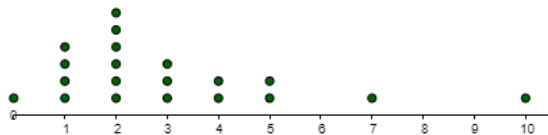
S

O

C

S

Practice FRQ. The dotplot shows the number of goals scored by the 20 teams playing in a city's high school soccer games on a particular day.



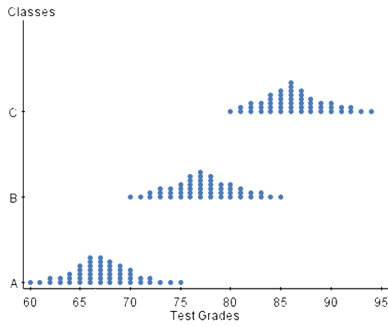
a) Describe the distribution.

b) One superstar scored six goals, but his team still lost. What are all possible final scores for that game? Explain.

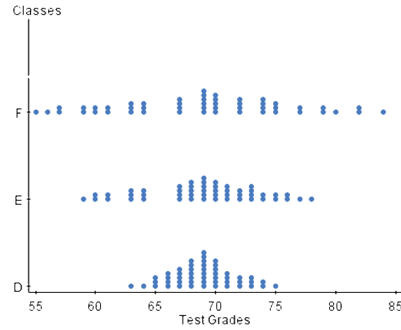
c) Is it possible that each of the teams that scored exactly two goals were the winners of that game? Explain.

Before moving on, check part (a): did you include context?

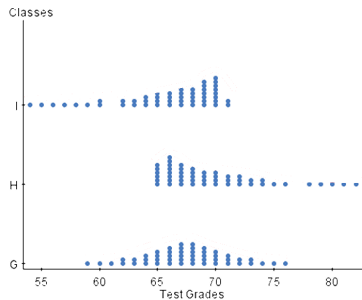
Comparing distributions warm-up. For each set of class score distributions, note the most distinctive difference:



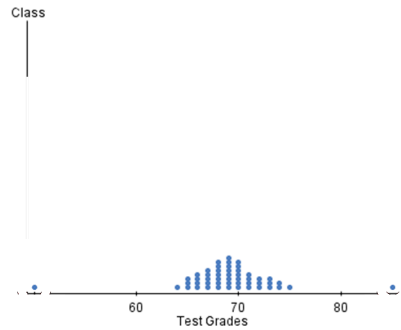
a) Classes A, B, C:



b) Classes D, E, F:

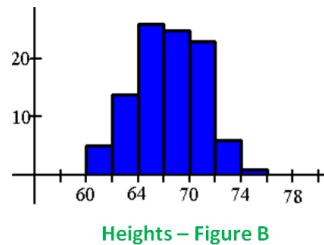
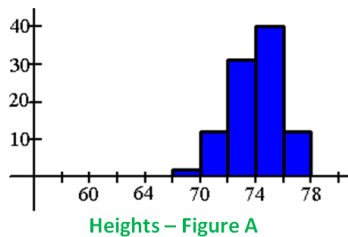


c) Classes G, H, I:



d) Class J:

If the Shoe Fits. Two histograms show heights of gymnasts (Figure A or B?) and female basketball players. Which is which, and which team gets the size-6 shoes? Why?



The **median** is the _____ of a distribution: it divides an ordered data set in half. If n is odd, the median is the middle value; if n is even, the median is the _____ of the two middle values.

Position of the median = _____

- Five fish are caught from the lake, lengths (in): 3 4 5 8 10. Median = _____
- A sixth fish is caught: 3 4 5 6 8 10. Median = _____
- Find the median: 8 4 2 1 9 2 3 7 5 4 8 6

The **mean** of a distribution of quantitative data is its _____ average. Notation: _____ is the mean of a sample; _____ (μ) is the mean of a population.

Find the mean length of the six fish (3, 4, 5, 6, 8, 10), then complete the deviation column:

VALUE	DEVIATION (VALUE – MEAN)
3	
4	
5	
6	
8	
10	
SUM	

The mean is the _____ point of the distribution: the positive and negative deviations always sum to _____.

Resistance. A measuring error makes the 10 in fish actually 15 in (then 20 in). What happens to each measure of center?

a) Median:

b) Mean:

A statistical measure is _____ if it isn't sensitive to extreme values.

Mean vs. median and shape. If a distribution is roughly symmetric with no outliers, the mean and median will be _____. If it is strongly skewed, the mean is pulled in the direction of the _____ but the median is not:

a) Skewed right: mean _____ median

b) Skewed left: mean _____ median

Challenge: Extensions of the Mean.

- a) A student's exam scores so far are 95, 83, 91, and 87. What does she need on the Unit 5 Exam to average 90 for the semester?

- b) Mr. Levy's 7th Period (31 students) averaged 3.2 points on the first knowledge check; 8th Period (18 students) averaged 4.1. What is the combined average? (The classes are not the same size!)

Position. You score 74% on the Unit 1 Exam. The class average was 66%, and you had the 2nd-highest score. Numbers without _____ are meaningless!

An individual's _____ is the percent of values in a distribution that are *less than or equal to* the individual's data value.

Jenny scored an 86 on her test. Use the class stemplot (25 scores) to find her percentile:

6 | 7

7 | 2 3 3 4

7 | 5 7 7 7 8 9 9

8 | 0 0 1 2 3 3 3 4

8 | 5 6 9

9 | 0 3

Key: 8|3 = 83 points

Which professor? Two brilliant professors teach the same course with the same curriculum, but they write their own final exams. Professor A's final always averages higher than Professor B's (85 vs. 82). Which professor do you choose, and why?

The **range** of a distribution is Maximum – _____. Caution: the range is a single _____, and it is *not* a resistant measure of variability.

Goals scored in 20 games by the 2016 U.S. women's soccer team: 5 5 1
10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3 Range = _____

The **standard deviation** measures the typical _____ of the values in a distribution from the _____.

How to calculate s_x :

- 1) Find the _____ of the distribution.
- 2) Calculate each _____ : value – mean.
- 3) _____ each of the deviations.
- 4) Add them all, then divide by _____. (This is the sample variance.)
- 5) Take the _____. This is the sample standard deviation.

Eleven students were asked how many close friends they have: 1 2 2 2 3 3 3 3 4 4

6. Work the calculation:

Properties: s_x is always $\geq$ _____ ; larger values mean _____ variation; s_x is *not* _____ ; and s_x measures variation about the _____ .

A **z-score** tells us how many _____ from the mean an observation falls, and in what _____ .

$z =$ _____

Jenny scored 86; the class mean is 80 with SD 6.07. Find and interpret her z-score:

Caution: do not say a z-score is a distance “away” from the mean; always say _____ (above) or _____ (below) the mean.

SAT or ACT? An admissions office is comparing two applicants, all else equal. One scored 1400 on the SAT; the other scored 32 on the ACT. Who should they admit? What would you need to know to decide fairly?

The **quartiles** divide an ordered data set into _____ groups of roughly equal size. Q_1 is the median of the values to the _____ of the median; Q_3 is the median of the values to the _____ .

$IQR =$ _____ – _____

Travel times for 20 New Yorkers (ordered, minutes): 5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85

Find the median, Q_1 , Q_3 , and IQR, then interpret the IQR in context:

Choosing measures of center and variability.

MEAN & SD

Always paired. _____.
resistant. Use when the data are
roughly _____ with no major
outliers.

MEDIAN & IQR

Always paired. _____. Use
when the data are _____ or
have major outliers.

Identifying outliers. Why look for them? They might be inaccurate values, they can indicate a remarkable occurrence, and they heavily influence non-resistant statistics.

Method 1: The $1.5 \times \text{IQR}$ Rule

Low outliers $< Q_1 - \rule{1cm}{0.4pt}$. High outliers $> Q_3 + \rule{1cm}{0.4pt}$

Method 2: Two Standard Deviations

An outlier is located more than _____ standard deviations above or below the _____ .

Highway fuel economy (mpg) for 25 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4 22.4
22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3 ($Q_1 = 22.2$,
 $Q_3 = 22.6$, $\text{IQR} = 0.4$)

Use Method 1 to find the outlier fences and identify any outliers:

Number of text messages sent yesterday by 70 students: $n = 70$, mean = 7.34, SD = 14.45, min = 0, $Q_1 = 2$, median = 3, $Q_3 = 7$, max = 104. The three largest values are 42, 42, and 104.

Use Method 2 to find the outlier boundaries and identify any outliers:

The **five-number summary**: _____, Q_1 , _____, Q_3 , _____.

A **boxplot** is a visual representation of the five-number summary.

Building a boxplot: draw the box from Q_1 to Q_3 , mark the median, extend whiskers to the last data values that are _____, and mark outliers as _____.

Draw a boxplot of the NYC travel times from 1D.10 (check for outliers first):

TRAVEL TIME (MINUTES)

Caution: boxplots do not display each individual value, and they don't show _____, _____, or peaks.

Transforming data converts observations to another scale. Effects:

ADDING / SUBTRACTING A CONSTANT A

Measures of center & location: _____ by a. Measures of variability: _____. Shape: _____.

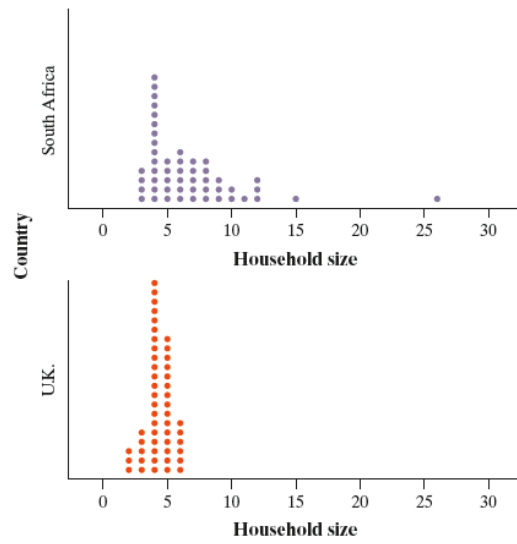
MULTIPLYING / DIVIDING BY B

Measures of center & location: _____ by b. Measures of variability: _____ by b. Shape: _____.

Australian students guessed their classroom's width in meters (actual: 13 m). The guesses: skewed right with two peaks, mean 16.02 m, median 15 m, SD 7.14 m, IQR 6 m. Define $\text{error} = \text{guess} - 13$, then convert the errors to feet ($\times 3.28$). For each transformation, give the new shape, center, and spread:

AP® Exam Tip. When comparing distributions, it is not enough to list values for each distribution's center and variability. You must explicitly _____ them, using words like "_____", "less than," or "about the same as."

Dotplots show household sizes for 50 randomly selected students from South Africa and from the U.K. Compare the distributions:



S

O

C

S

Two non-negotiables in every comparison: context, and comparative language.